EXHIBIT G

DOE/NIH Human Genome Contractors/Grantee Workshop

November 3-4, 1989

Santa Fe Institute

Santa Fe, NM

An Expanding Cosmid Contig Map for Chromosome 19.

L. Ashworth, E. Branscomb, L. Brown, C. Chen, P. de Jong, A. Fertitta, E. Garcia, J. Garnes, J. Lamerdin, F. Lohman, H. Mohrenweiser, D. Nelson, W. Nelson, A. Olsen, B. Perry, T. Slezak, K. Tynan, M. Wagner, P. Wilkie and A.V. Carrano. Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA.

High resolution gel electrophoresis provides a simple and versatile method for DNA fingerprinting and the creation of contigs or sets of overlapping genomic clones. Cosmid libraries are constructed from YAC clones or from flow–sorted chromosomes. Cosmid DNA is isolated by an alkaline–lysis procedure and each cosmid is digested with EcoRI to measure insert size and concentration. The isolated DNA is then cut with a combination of five restriction enzymes and the fragment ends labeled with one of four different fluorochromes. Our approach to contig construction: 1) uses a robotic system to label restriction fragments from cosmids with fluorochromes; 2) uses an automated DNA sequencer to capture fragment mobility data in a multiplex mode (i.e. three cosmids and a size standard in each lane); 3) processes the mobility data to determine fragment length and provide a statistical measure of overlap among cosmids; and 4) displays the contigs and underlying cosmids for operator interaction and access to a database. We have applied these methods to construct a cosmid contig map for a 600 kbp YAC clone from chromosome 14 and are currently analyzing cosmids to construct contigs for all of chromosome 19. Throughput rate is currently about 48 cosmids per day per machine but 96 cosmids per day is achievable. Resolution of fragment size is to within 1–1.5 bases over the range of 29–462 bases for which data are captured. The more than 2500 chromosome 19 cosmids analyzed to date assemble into over 320 contigs with an average contig length of 3.2 cosmids. Many of these contigs have been located to the chromosome by fluorescence *in situ* hybridization and also mapped to known genes. The "minimal" spanning sets of cosmids provide unique starting material for genome sequencing. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W–7405–ENG–48.

P01

PROGRESS OF "TOP-DOWN" MAPPING APPROACHES TO CHROMOSOME X

David F. Barker, Arnold R. Oliphant, Pamela R. Fain, David E. Goldgar, Huntington F. Willard, Anne Vincent, Stephen Warren, Jennifer Puck, Robert L. Nussbaum and Christine Petit

The initial focus of work aimed at defining the genetic and molecular structure of the X chromosome has been the isolation and ordering of a set of clones defining RFLP markers. We have isolated more than 80 such markers for the X utilizing an X library, LAOXNL01, constructed by LANL under the auspices of the DOE human genome project. If combined with the over 100 other X RFLP markers that have been isolated in many laboratories, these markers would comprise a set with an average spacing better than 1 per megabase. The complete and reliable ordering of such a set would provide a useful "backbone" structure for both high resolution genetic-disease mapping and the ordering of sets of overlapping clones, "contigs", with respect to each other.

The ordering approaches we have used include genetic mapping in the CEPH linkage reference families and in genetic disease families. We have also tested a variety of human-rodent hybrid cell lines containing unique segments of the X chromosome to provide additional ordering information. The physical breakpoints have been generated by natural translocations or deletions or by strategies designed to select broken chromosomes in tissue culture lines. The latter include "pushmi-pullyu" hybrids generated by Brown et al. (HGM10) and "radiation hybrids" isolated as described by Cox et al. (AJHG 43: A141). The physical breakpoints characterized to date divide the chromosome into 25 regions, 8 on Xp and 17 on Xq. Nearly all of the 80 polymorphic markers which we have isolated have been mapped into one of these 25 regions and a summary map will be presented. The current status of the genetic linkage map will also be shown.

TOWARDS CONSTRUCTION OF A 2Mb PHYSICAL MAP OF HUMAN CHROMOSOME 16.

D.F. CALLEN, V.J. HYLAND, L.Z. CHEN, J.C. MULLEY, S. LANE, E.G. BAKER, G.R. SUTHERLAND

Cytogenetics Unit, Adelaide Children's Hospital, North Adelaide, South Australia 5006.

The use of mouse/human hybrids containing portions of human chromosome 16 provides a method for rapid physical mapping. We aim to extend our hybrid panel of this chromosome until the chromosome can be subdivided into approximately 50 intervals. At this point the average interval spanned by breakpoints will be 2Mb.

The human parent used in the construction of these hybrids are chromosome 16 translocations and deletions which have been identified as the result of cytogenetic investigations. These have been obtained from our laboratory, from the NIGMS cell repository and from the kind co-operation of many other cytogenetic laboratories. Each newly generated hybrid containing a derived human chromosome 16 is evaluated using a battery of probes which have been already physically mapped to chromosome 16. This allows the ordering of the breakpoints of the new hybrid in relation to the existing physical map.

In conjunction with the generation and evaluation of new hybrids further probes on chromosome 16 will be physically mapped. These include gene probes, probes that we have generated from a lambda library derived from the hybrid CY3, single-copy probes on chromosome 16 obtained from Dr. P. Harris, chromosome 16 cosmid clones from Dr. M. Breuning, and probes which have been genetically mapped in the CEPH pedigrees from Dr. C. Julier. It is planned to map other polymorphic probes which have been genetically mapped from Dr. P. O'Connell and to locate cosmid contigs in collaboration with Dr. E. Hildebrand. This will lead to a detailed correlation of the genetical and physical maps of chromosome 16.

At the present time we have constructed twenty hybrids of chromosome 16 which, with the use of the three rare fragile sites on this chromosome, can subdivide the chromosome to potentially 24 intervals. At the present time 83 protein markers, gene probes and anonymous DNA fragments have been physically napped to all, or a subset of, these hybrids.

This hybrid panel can provide a useful resource for work involving other chromosomes. Breakpoints on chromosomes 1, 3, 4, 9, 10, 11, 12, 13 and 22 are included in the panel. All hybrids are available from Dr. David F. Callen on request.

This work is supported by the Department of Energy Grant DE-FG02-89 ER60863. This support does not constitute an endorsement by DOE of the views expressed in this abstract.

Generation of a Physical Map of the Long Arm of Human Chromosome 11.

G.G. Hermanson*, P. Lichter#, D.C. Ward#, and G.A. Evans*.
*Molecular Genetics Laboratory, The Salk Institute, La Jolla, CA 92138
#Department of Human Genetics, Yale University School of Medicine,
New Haven, CT 06510

Many loci implicated in human diseases have been mapped to the long arm of human chromosome 11 including: ataxia telangiectasia, tuberous sclerosis, multiple endocrine neoplasia type 1, and translocations found in Ewing's sarcoma, and acute leukemias. This region also contains genes which are members of the immunoglobulin superfamily: NCAM, CD3 γ, δ, and ε, and Thy-1, as well as the oncogene c-ets-1 and the leukocyte marker CD5. Given that only a small proportion of human genes have been identified and mapped, it is clear that many more genes which might be important in human development or disease may be contained in this region. To generate a physical map and localize these potentially important genes, we have isolated linking clones containing multiple rare-cutting restriction enzyme sites. These clones can be used as probes for pulsed-field gel and fluorescent in situ hybridization techniques to physically map this region, and identify potential HTF islands that are associated with many gene coding regions. In order to identify these linking/HTF island clones, a cosmid library was constructed from a somatic cell hybrid line containing only the long arm of human chromosome 11 from 11q12-11qter. Cosmid clones containing human insert DNA were selected by hybridization to total human genomic DNA and picked into a total of ten 96-well plates for further analysis. A Beckman robotic workstation was used to prepare miniprep DNA from this cosmid collection, as well as to assay the 960 clones for the presence of the rare cutting restriction enzyme sites Not I, Sac II, BssH II, Pvu I, Mlu I, and Sfi I. A total of 175 cosmid clones contained at least one Not I site in their insert. Since Not I sites are rare in the genome, cosmids were initially selected for further analysis only if they contained at least one of these sites. Thirty-two unique cosmids were finally selected that contain at least one Not I site and sites for the majority of the other rare cutting enzymes. These 32 linking/HTF island cosmids, as well as cosmids containing previously identified genes are being ordered into a molecular and cytogenetic map of chromosome 11 by single copy fluorescent in situ hybridization and pulsed-field gel electrophoresis techniques.

CONSTRUCTION OF HUMAN CHROMOSOME 21 SPECIFIC YEAST CHROMOSOMES

Mary Kay McCormick[1,2], James H. Shero[4], Mei Chi Chung[3], Yuet Wai Kan[3], Philip Hieter[2,4], Stylianos E. Antonarakis[1,2]

[1]Genetics Unit, Department of Pediatrics, [2]Predoctoral Training Program in Human Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205
[3]Howard Hughes Medical Institute, University of California, San Francisco, CA 94143
[4]Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205

Chromosome 21 specific yeast artificial chromosomes (YACs) have been constructed by a method that performs all steps in agarose, allowing size selection by pulsed field gel electrophoresis and the use of nanogram to microgram quantities of DNA. The DNA sources were hybrid cell line WAV-17, containing chromosome 21 as the only human chromosome, and flow sorted chromosome 21. The transformation efficiency of ligation products was similar to that obtained in aqueous transformations and yielded YACs with sizes ranging from 100 kilobases to over 1 megabase when polyamines were included in the transformation procedure. Twenty five YACs containing human DNA have been obtained from the mouse-human hybrid, ranging in size from 200 kb to > 1000 kb with an average size of 410 kb. Ten of these YACs were localized to sub regions of chromosome 21 by hybridization of riboprobes (corresponding to the YAC ends recovered in E. Coli) to a panel of somatic cell hybrid DNAs. Twenty one human YACs, ranging in size from 100 kb to 500 kb with an average size of 150 kb, were obtained from ~50ng of flow sorted chromosome 21 DNA. Three were localized to subregions of chromosome 21. Yeast artificial chromosomes will aid the construction of a physical map of human chromosome 21 and the study of disorders associated with chromosome 21 such as Alzheimer's disease and Down syndrome.

Assembly of Cosmid Contigs in the Region of the Nucleotide Excision Repair Genes on Human Chromosome 19.

Mohrenweiser, H. W., de Jong, P. J., Perry, B. A., Tynan, K. T., Lohman, F. P. and Carrano, A. V. Biomedical Sciences Division, L-452, Lawrence Livermore National Laboratory, Livermore, CA 94550

The genes CKMM, ERCC1 and ERCC2 have been assigned to the region q13.2 - q13.3 of human chromosome 19 and have been localized to within ~250 kb of each other by PFG electrophoresis. A human chromosome 19 specific cosmid library was screened with a pool of probes for these genes. The DNA from each of the selected cosmids was analyzed using a high density restriction enzyme site mapping ("fingerprinting") strategy (Carrano et al., Genomics 4:129, 1989). Overlapping cosmids were detected, and contigs assembled, based upon commonality of restriction enzyme digest fragment sizes. Two contigs could be generated from the fingerprinting data obtained from analysis of the group of cosmids initially selected by probing. Upon reprobing of this group of selected cosmids with the individual probes, the 2 cosmids in one contig hybridized with the ERCC1 probe while the second contig of 3 cosmids contained the CKMM gene. Four additional cosmids, analyzed during the fingerprinting of ~2200 random cosmids (~1.4X library), have been linked to the most centromeric cosmid of the original CKMM contig yielding a contig with a tiling path of 6 cosmids. No cosmids containing the ERCC2 gene were isolated and no randomly selected cosmids have been linked to the most telomeric CKMM cosmid, thus a gap of ~10kb exists between the CKMM contig and a previously isolated set of cosmids containing the ERCC2 gene. The ERCC2 gene is ~150kb from the ERCC1 contig. An expressed sequence overlapping the ERCC1 gene is within this region, thus at least 4 expressed genes, comprising >60kb, are within this 250kb region. Closure of the gaps, as necessary to form a single contig containing the ERCC1, ERCC2 and CKMM genes should be attained with very limited walking, although it will apparently be necessary to screen additional libraries (YAC, lambda, cosmid) to complete the contig.

Work performed under auspices of the US DOE by the Lawrence Livermore National Laboratory; contract No.W-7405-ENG-48

Contig Assembly and Characterization of a Chromosome 19q Specific Minisatellite Element.

Tynan, K. T., Mohrenweiser, H. W., Branscomb, E. W., deJong, P. J. and Carrano, A. V.
Biomedical Sciences Division, L-452, Lawrence Livermore National Laboratory, Livermore, CA 94550

A minisatellite consisting of a repeat of six 37 bp elements was identified by Das et al. (J Biol Chem 262: 4787, 1987) in intron six of the ApoCII gene, a gene mapping to 19q13.2. Five elements of this minisatellite are present in the ERCC2 locus, another locus on chromosome 19q13.2 (Weber et al., EMBO J in press). It has been estimated that this minisatellite exists at ~60 "loci." Additional evidence suggests this minisatellite is localized to the q13 region of chromosome 19. Approximately 150 cosmids were identified as containing this minisatellite during the screening of a human chromosome 19-specific cosmid library (~7000 cosmids) with a probe containing the repeat element. Sixty seven of these cosmids have been analyzed by fingerprinting (Carrano et al., Genomics 4:129, 1989) during the analysis of ~2200 random cosmids. Twenty six of the 67 cosmids have been assigned as members of 16 contigs. Three of these contigs are comprised of only 2 cosmids, both members of each pair being minisatellite positive. Of the remaining 14 contigs, the minisatellite is present as 3 adjacent cosmids in 3 contigs, a pair of overlapping cosmids in one additional contig and only once in the 9 other contigs. As expected, this appears to be an efficient strategy for establishing "seed" contigs in this region of chromosome 19 and also assists in validating the contig building strategy. One of the minisatellite containing cosmids is within the PVS locus contig at 19q13.2, thus at least 3 of the minisatellite repeat "loci" are located within functional genes. A number of additional minisatellite containing cosmids have been mapped by *in situ* hybridization and all map to the 19q13.2-q13.4 region. Hybridization analysis of DNA from a human lymphoblastoid cell line and hamster cells containing all or parts of human chromosome 19 following either normal or PFG electrophoresis are consistent with the previous estimate of ~60 loci and localization of these elements to human chromosome 19q.

Work performed under auspices of the US DOE by the Lawrence Livermore National Laboratory; contract No.W-7405-ENG-48

Abstract for DOE Contractor-Grantee Workshop November 3-4, 1989

CHARACTERIZATION OF HUMAN CHROMOSOME-SPECIFIC PARTIAL DIGEST LIBRARIES IN LAMBDA AND COSMID VECTORS.

Kathy Yokobata, Jennifer McNinch, Lee Pederson, Marvin A. Van Dilla and Pieter J. de Jong, Biomedical Sciences Division, Lawrence Livermore National Laboratory, P. O. Box 5507, Livermore, CA 94550

As part of the National Gene Library Project we have constructed partial digest chromosome-specific human genomic libraries for studies of genetic disease, physical mapping of chromosomes, and other studies of the molecular biology of genes and chromosomes. New procedures were developed for isolating DNA of high molecular weight from flow sorted human chromosomes and for preparing large insert libraries in lambda replacement and cosmid vectors from small quantities (about 1 µg) of DNA. These procedures have been used successfully for the preparation of lambda and cosmid libraries specific for chromosomes 19, 21, 22 and Y. A large insert lambda library has been prepared from sorted chromosome 11 DNA as well. The lambda vectors used were Charon 40 and GEM 11; the cosmid vector was Lawrist 5, a lambda origin cosmid vector.

For all libraries, we have estimated purity by flow karyotype analysis. We have also examined the purity of the lambda libraries by plaque hybridization and have examined the origin of clones which show no signal in these plaque hybridizations. The average insert size was determined by excising insert DNA from randomly selected clones for each lambda library. In addition, we have screened the chromosome-19 lambda library with probes known to map to the chromosome to establish representation of these sequences in the library.

We have begun characterization of the cosmid libraries by mapping randomly selected cosmid clones by fluorescent in situ hybridization. We have examined the stability of insert sequences in cosmid libraries in various bacterial hosts and describe studies with an unstable cosmid clone from the CKM locus on chromosome 19 and with insert sequences in the chromosome-Y library. Much of the characterization of the chromosome-19 cosmid library is being done under the auspices of the chromosome 19 ordering project.

The libraries for chromosomes 11 and 19 were made from chromosomes sorted from monochromosomal hybrid lines, whereas the libraries for 21, 22 and Y used human cell lines for the starting material. Because of the purity advantages of obtaining sorted chromosomes from monochromosomal hybrids and the new availability of sortable hybrids for these chromosomes, new 21, 22 and Y libraries are being constructed. We are currently isolating chromosomal DNA from chromosomes 3 and 12 for lambda and cosmid libraries.

Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract number W-7405-ENG-48.

P08

Lawrence Livermore National Laboratory

October 12, 1989

Dr. Sylvia Spengler
LBL - Human Genome Center
459 Donner Laboratory
Berkeley, CA 94720

Dear Dr. Spengler:

The following is in response to your letter dated September 26, 1989. It is an abstract for the poster session at the DOE Human Genome Project Contractor/Grantee workshop.

Title: A Natural Language Query System for Genbank

Authors: Michael Cinkosky (LANL), Rowland R. Johnson (LLNL) and Rob Pecherer (LANL)

Abstract:

The Genbank database contains a wide variety of information that is of interest to researchers involved in the Human Genome Project. The logical structure of the database is complex enough that procedures to query the database are required. Such procedures have been implemented or will be implemented for the most typical classes of queries. Atypical queries require a procedure to be constructed by the end user. It is unlikely that a Human Genome Project researcher will be able to construct a procedure that will satisfy an atypical query. Thus, the usefulness of Genbank to the Human Genome Project will be restricted.

A possible solution to this problem is a system that will accept a query stated in English and translate it to a procedure appropriate for the database system. Towards this end, a prototype of such a system was developed in July 1989 for the Human Genome Project at LANL. This system is based on a commercially available product that translates English into a sequence of relational database query commands.

The work undertaken in the development of the prototype was to customize the product to work in the Genbank domain. This customization is a continuing, but diminishing, effort. New users of the system will bring out unrealized English constructions that must be incorporated. As more and more people use the system, there will be fewer unrealized English constructions to incorporate.

Currently, the prototype contains a subset of the data that will be in Genbank. The system will be available at the conference so that participants may try queries on the system.

P09

Title: GnomeView: A Graphics Interface to the Human Genome

Authors: Richard J. Douthart, David A. Thurman and Victor B. Lortz

Abstract:

Pacific Northwest Laboratory is developing GnomeView, a software system that provides a graphical interface to the large quantities of data and information generated by the Human Genome Initiative. GnomeView allows the user to visually browse and manipulate color graphic representations of genetic maps, physical maps, and sequences. These representations provide the user with a sense of topology and reveal patterns in the data that are otherwise difficult to detect. This hierarchy of mappings can be traversed using the pan-and-zoom capabilities of GnomeView and all objects in maps will be queryable.

GnomeView uses the X Window System to render and display maps and sequences on a UNIX workstation. It is being developed on a Sun workstation in portable C and should be portable to any UNIX workstation. It includes db_VISTA, a network-model database, that permits natural and efficient representation of the relationships in genomic information. Landmarks, features, and blocks of sequence are stored in linked lists of objects in the database. Objects common to more that one mapping need only be stored once in the database, but may be referenced in many mappings. Specific attention has been paid to designing database algorithms and user interface techniques that scale well to high data volumes.

GnomeView accepts queries from the user and responds in a number of different ways. Depending on the query, these responses can take the form of maps, textual lists, or color-coded histograms. Further information can be obtained by querying these responses. As an example, the hierarchy of the superoxide dismutase loci on Chromosome 21, from band location, to restriction map, to base sequence, will be presented.

Identification of genes in anonymous DNA sequences

C. A. Fields and C. A. Soderlund

Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001, USA. 505-646-5466.

The objective of this project is the development of practical software to automate the identification of genes in anonymous DNA sequences from the human, and other higher eukaryotic genomes. A prototype automated sequence analysis system, **gm**, has been implemented in C to run on Unix workstations. This system accepts as input: i) a DNA sequence, ii) consensus matrices for locating splice sites, translational start sites, and polyadenylation sites, iii) match-quality cutoffs for consensus searches, and iv) base frequency and codon usage standards for coding regions and introns. It produces as output both schematic models of possible genes contained in the sequence that show the locations of the coding sequences, introns, and control signals, and predicted amino-acid sequences for each of these possible genes. The models include the numerical results of evaluating each of the component exons and introns.

gm has been extensively tested on *C. elegans* sequences in the 10 kb size range containing known genes of up to 10 exons, and is capable of generating complete, correct analyses showing all possible alternative splicing patterns. Such analyses typically require a few minutes running time on a Sun 4/60 workstation, depending on the stringency of the search parameters used. Current effort is focussed on improving the pattern recognition and statistical analysis modules used by **gm**, and on implementing greedy algorithms for performing fast first-pass analyses using low stringency parameters.

Sequence Matching and Motif Identification Abstract

Daniel Gusfield
Computer Science Division, U.C. Davis *and*
ICS Division, Lawrence Berkeley Laboratory

Eugene L. Lawler William I. Chang
Computer Science Division, U.C. Berkeley *and*
ICS Division, Lawrence Berkeley Laboratory

Frank Olken
ICS Division, Lawrence Berkeley Laboratory

Sequence Matching and Alignment

Dynamic Programming has been the technique of choice for sequence matching problems arising in biology. However, DP algorithms are likely to be impracticable for problems of the size that will be created by the massive amounts of data generated by the Human Genome Initiative. On the other hand, there are other well developed techniques based on the use of suffix trees or finite state machines, which are more efficient than DP for many sequence matching problems. Our work is focussed on extending the use of suffix trees to additional problems in computational biology, with the objective of developing computational methods that are faster and more practicable than those which currently exist.

The principal difficulty in using suffix trees in biological applications is that they are best adapted for problems in exact, rather than approximate, sequence matching. However, we have found ways to use suffix trees to improve existing efficient algorithms for certain approximate sequence matching problems. For example, we have been able to use suffix trees to replace hashing in the Lipman-Pearson algorithms. By first building a suffix tree of the shorter sequence we can, in a single left-to-right scan of the longer sequence, compute for each position the longest exact match with any portion of the shorter sequence. This yields information which encompasses the Lipman-Pearson hashing step for all choices of tuple size simultaneously. This subroutine also enables us to simplify and make practicable a theoretical result of Landau and Vishkin, namely that matching up to k indel/substitution errors can be accomplished in time a factor k worse than linear. Our results further indicate that when the error threshold is below 15 percent for nucleotides (in particular errors arising from sequencing) almost all mismatches can be eliminated from consideration very quickly so that in linear time we can expect to find every match. Applications include the problem of computing overlaps in sequence assembly when sequencing errors are not negligible.

Another area of interest to us is that of sensitivity analysis. Existing algorithms for approximate matching require the user to specify mismatch (indel/substitution) penalties. However, these penalties are not known exactly. We have already had some success in developing efficient parametric algorithms which reveal the optimal sequence matching solution as a function of the penalty costs.

Searching for Motif Patterns

It is often possible to develop very efficient special purpose algorithms that search for repeat patterns in a sequence. Examples include the identification of *inverted* and/or *complemented* repeats with gaps (there are no a priori upper or lower bounds placed on gap lengths). For concreteness, a pair $..A.....A'..$ is a *maximal inverted pair* if A' is an inverted copy of A and the sequence does not look like $..AB..B'A'..$ or $.BA.....A'B'$. Note that the same subsequence can be part of more than one maximal inverted pair: in the sequence ..abc.xba.cba.., both ..abc.....cba.. and ..ab..ba..... may be maximal. The problem of identifying all such maximal pairs (including positions) can be solved in quadratic time and space by dynamic programming methods. However, we have recently developed a very simple method which finds all maximal inverted pairs in linear space and in time proportional to the length of the sequence plus the number of such pairs. Importantly, our method can be modified to find only pairs above a certain length.

DNA Bend Detection

Marjorie S. Hutchinson *
Information and Computing Science Division
Lawrence Berkeley Laboratory
Berkeley, Ca, 94720

We will describe and demonstrate a program on a Sun workstation that examines DNA sequences for bends or kinks. The program has the capability of scanning either a single sequence or a specified portion of Genbank to identify likely candidates for a bend. We have developed a user-friendly interface that allows the user to examine a candidate or list of candidates and view the parameters that indicate whether a bend might be present. Some of the more important output parameters are plotted against residue number and these plots are analyzed for the likelihood of a bend. Threshold values for the various tests for a bend can be set by the user. We display the calculated 3D structure of the sequence, allowing the user to rotate the structure, and zoom in on interesting portions.

The structural properties of the DNA sequence are derived using software provided by Wilma Olson of Rutgers University. [1] Her program utilizes the potential energies of interaction of free base pairs (as calculated by Srinivasan et al.) [2] to calculate a static structure. This structure is determined by choosing the local conformational geometries which optimize the computed average orientations of adjacent residues.

The Olson program also calculates several measures of chain stiffness including the persistence length, which is a measure of the distance over which the initial direction of the DNA is preserved. For each residue in the chain, it also calculates: the average angular orientations of the sequence, the average vectorial displacement, and the mean-square end-to-end distance. Variations in these values over the chain length may indicate with bends. We display this data and utilize Fourier transform coefficients of the resulting curves as further evidence for or against bending.

Execution of the program in its current stage of development will be demonstrated and planned improvements will be described.

*email: margeh@csam.lbl.gov

[1] W.K. Olson and A.R. Srinivasan, (1988), The translation of DNA primary base sequence into three-dimensional structure. *Cabios*, 4, 133-142.

[2] A.R. Srinivasan, R. Torres, W. Clark and W.K. Olson. Base sequence effects in double helical DNA. I. Potential energy estimates of local base morphology, (1987) *J. Biomol. Struct. Dynam.*, 5, 459-496.

Robotic Control of the Laboratory Procedure for Clone Candidate Selection

S. Lewis, J. Gingrich, and J.C. Bartley
Engineering and Life Sciences Divisions
Lawrence Berkeley Lab
Berkeley, California

We have automated a laboratory procedure to detect and select potentially transformed yeast cultures, by adapting the standard laboratory technique to a standard 96 well microtiter plate format and by programming a general purpose robot to carry out the procedure. The robot runs unattended and can screen up to 960 candidates in less than an hour.

The procedure steps are as follows:

- Initial transformed spheroplasts which have been grown up suspended in solid media are picked by hand into selective liquid media in the wells of microtiter plates.
- These microtiter plates are placed into the robot's incubator station and the robot application is started.
- After incubation the robot transfers each plate in turn to a plate reader. Growth is determined according to the turbidity in each well.
- For each well in which there is growth the robot pipets an aliquot into a new well of selective media, which the robot has previously filled. Only those wells in which growth has occurred are transferred, resulting in fewer plates.
- These plates are then placed in the incubator by the robot for subsequent incubation.

Data describing the identities and characteristics of the selected cultures are recorded for subsequent inclusion in a laboratory notebook database. We also discuss specific problems that arose automating this procedure such as: selection of well bottom shape (flat, u, or v), or suspension and aeration of the yeast growing in liquid media. A video tape of the procedure will illustrate the robot's capabilities and each individual outlined above.

ALGORITHM FOR SEQUENCE GENERATION FROM K-TUPLE WORDS

CONTENT: Labat, I., Drmanac, R., Crkvenjakov,R. Genetic Engineering Center, PO Box 794, 11000 Belgrade, Yugoslavia

Any text as well as nucleotide sequence, can be represented as a set of the overlaping k-tuple words, similary to the methods applied in the most efficient sequence comparison algorithms used today. Our algorithm uses intrinsic nucleotide sequence informatics to regenerate the original sequence without k-tuple position and frequency information inherent to the former algorithms. K-tuples are ordered by maximal overlaping up to the moment when none, or two or more k-tuples overlap with the last one attached. Further ordering is ambiguous. A primary subfragment (SF') is thus defined. Number of the heuristic methods developed repairs and unambiguously connects SF' into the real subfragments (SF). Number and length dispersion of the SFs as sequence informatics entities depends on length and simplicity of the sequence as well as length of k-tuples and extent of mistakes in the set. The other part of the algorithm enables regeneration of the sequence of the length of the human genome fragmented in the suggested manner (Drmanac et al., GENOMICS 4 , '89, 114). It consists of several k-tuples sets and SFs manipulations on different, overlaping sequence fragments. Our software is applied on the IBM PC/AT compatible. In simulation experiment on the 50kb sequence, complete k-tuples sets (k=8 to 12, depending on GC content) of the consecutive nucleotide sequence fragments up to 900 bp were handled. In over 91% of analyzed fragments the complete sequences were regenerated. In remaining cases, sequences were regenerated to the level of several (below 15) SFs. Also, 10% of false negative k-tuples in the set makes no problem in most of analysed sequences. Further improvements of algorithm are needed (and suggested) for complete regeneration of some specific sequences, and for using sets with more false positive and false negative k-tuples.

Automated Extraction of Band Data from Digitized Autoradiogram Images

S. Lewis, K. Gong, J. Jaklevic, W. Johnston, E. Theil
Engineering Division
Lawrence Berkeley Lab
Berkeley, California

We are developing computer methods to analyze electrophoresis gel and autoradiogram images generated in a production mapping laboratory. The objective is to distill from the digitized images essential band size and intensity information. As electrophoresis gels display a lot of operational differences, the analysis software must handle problems such as: diffuse or sharp bands, overlapping bands, crooked lanes, over and under exposure times and varying DNA migration rates. Though the program provides fully automatic analysis of the bands if desired, there is the opportunity at every step for the operator to substitute his own judgement. For instance, the correct position of crooked lanes is indicated by the operator pointing and clicking the length of the lane. By isolating each component step in the analysis the flexibility of substituting either user results or improved algorithms is easily achieved.

The filtering techniques which we are using to detect bands are described. Currently this is a simple single pass non-adaptive digital filter which removes the background and slopes in the original one-dimensional data. The filtered 1-D data is then analyzed for peaks indicating bands. Once the bands have been located the sizes are calibrated according to their relative position within the lane.

The sets of band data thus derived is subsequently used for algorithmic comparisons between lanes and gels, as well as in mapping algorithms. The band data and pointers to the original archive images are stored in a laboratory data base.

ImageQuery: An Interface to a Biological Images and Videos Database

S. Lewis and W. Johnston —
Engineering and Computer Science Divisions
Lawrence Berkeley Lab
B. Morgan and S. Jacobson
Advanced Technology Planning
University of California at Berkeley
Berkeley, California

ImageQuery is a software tool which combines textual and visual methods for organizing and querying an image database. It was developed by the Advanced Technology Planning group at UC Berkeley to provide online access to collections of primarily visual materials. One of the initial applications at UCB provided information on archeological artifacts. In our case it has been adapted to catalog autoradiograms.

The ImageQuery interface enables researchers to search through images based upon either selectable fields of descriptive information using boolean logic, or to browse through iconic representations of the images themselves. ImageQuery icons represent the type of image and a unique identifier.

Once a set of images has been selected from the database via any combination of selection methods, ImageQuery can invoke specific analysis programs directly. For example, the program AnGel is used to locate and determine the sizes of bands within autoradiograms, or the program HIPsTools which provides users with zoom, pan, distance measurements, and other standard image processing capabilities.

ImageQuery runs on Sun work stations and can be interfaced either to a flat file of descriptive parameters, or to Ingres, a relational database management system. A demonstration of the query capabilities will be given using selected electrophoresis gel image data.

Supercomputer Simulations and Experimental DNA Electrophoresis*

Lim, H.A., Burnette, D.E., and McMullen, D.F.

Supercomputer Computations Research Institute

Florida State University

Tallahassee, FL 32306-4052

The main objective of this project is the development and optimization of general purpose supercomputer algorithms so that the mobility of large DNA molecules in electrophoresis can be simulated. By working in parallel with laboratory experimentalists, this project should integrate new supercomputer-based DNA electrophoresis simulating programs with innovations in electrophoretic techniques, and biochemical manipulations of chromosomal DNA. Though electrophoresis has been sucessfully used to separate chromosomes from lower eucaryotes (*e.g. Saccharomyces cerevisiae*) and other simple organisms (*e.g. Drosophila melanogaster*), the amount of DNA in the average human chromosome (about 1.4×10^8 bp) is at least 10-fold to 100-fold greater than current record size separable by electrophoresis. The project will focus on three major aspects of the problems associated with the current state of the art of electrophoresis: (1) Size—the current size record separable is still at least a factor of 10 smaller than the average size of human chromosome; (2) Speed—the current rate of data collection and analysis can still be improved; and (3) Resolution—errors in DNA sequence determination occur most commonly by insufficient resolution of DNA fragments in electrophoresis due to band inversion or compression. Unless this can be overcome, the human genome sequence determined by electrophoresis will be of little use.

* Funded by DOE Energy Research through SCRI

Human Genome Management Information System

Betty K. Mansfield, Judy M. Wyrick, John S. Wassom, Po-Yung Lu, Mary A. Gillespie, and Sandy E. McNeill
Health and Safety Research Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6050
(615) 576-6669, FTS 626-6669

The Human Genome Management Information System (HGMIS), sponsored by the U.S. Department of Energy (DOE) at the Oak Ridge National Laboratory, has the following roles in the Human Genome Program: (1) to assist the DOE Office of Health and Environmental Research (OHER) in communicating issues relevant to human genome research to DOE contractors and grantees and to the public and (2) to provide a forum for exchange of information among individuals involved in genome research or the development of instrumentation and methodologies to implement genome research. To fulfill these communications goals, HGMIS is producing technical reports, DOE Human Genome Program reports, a quarterly newsletter, and an electronic bulletin board. These documents/facilities are available to all interested persons upon request. The first technical report will assess instrumentation and methodology development relevant to DNA mapping and sequencing. The DOE Human Genome Program reports will contain information on human genome research and development activities supported by the DOE program as well as background information. The *Human Genome Quarterly* newsletter features technical articles, meeting reports, a calendar of genome events, announcements, and other information relevant to genome research. Accessible via modem through direct dial or via the user's host mainframe computer network, the electronic bulletin board contains information organized by categories (e.g., menu, general information, news and comments from OHER; summaries and highlights of research projects; meeting announcements/calendar; literature highlights; DOE Human Genome Program contacts; and international activities). HGMIS welcomes comments, suggestions, and contributions from the genome research community.

SDT : A DATABASE SCHEMA DESIGN TOOL *

Victor M. Markowitz and Frank Olken
Computer Science Research and Development Department
Lawrence Berkeley Laboratory
1 Cyclotron Road, Berkeley, CA 94720

We present a database schema design tool (*SDT*) developed at Lawrence Berkeley Laboratory. The purpose of *SDT* is to provide a powerful and easy to use design interface for biologists, and to increase the productivity of the database design process. This entails insulating the schema designer from the underlying database management system (DBMS).

For the schema design interface we have chosen a version of the *Data-Flow* (DF) model for describing processes and process correlations, and a version of the *Extended Entity-Relationship* (EER) model for the specification of the static structure of information systems. The EER model we use includes, in addition to the basic construct of object (entity and relationship), both generalization and full aggregation abstraction capabilities. We have developed an integrated DF/EER schema design methodology. Following this methodology, DF specifications are represented by EER constructs, so that design of an information system results in an EER schema that captures both the structural and functional characteristics of the modeled system. Once an EER schema is specified, *SDT* is employed in order to generate the corresponding DBMS schema.

SDT consists of two main modules, SDT_R and SDT_{DM}. The first module, SDT_R, takes EER schemas as input and generates abstract relational schemas. SDT_R consists of three parts: the canonical mapping of EER schemas into normalized relational schemas; the assignment of names to relational attributes; and merging relations. The canonical mapping generates relational schemas, including key and referential integrity constraints. The high normal form (BCNF) of this schema ensures efficient update performance by the DBMS. Name assignment can be customized in order to meet the needs of the user (e.g. short names, minimum number of attributes, etc.). Finally, merging of relations reduces the number of relations, thus improving query performance.

The second module, SDT_{DM}, takes abstract relational schemas as input and generates relational DBMS (e.g. SYBASE, DB2, INGRES) schemas. For a DBMS that supports the specification of *triggers*, such as SYBASE, the main part of STD_{DM} consists of generating the appropriate *insert*, *delete*, and *update* triggers corresponding to the referential integrities associated with the abstract relational schema.

The research related to *SDT* is presented in [1] and [2]. The DF/EER design methodology is described in [3] and an example application is presented in [4]. The logical algorithms of *SDT* and their implementation are described in [5]; *SDT* was implemented using C, LEX, and YACC, on Sun 3 under Sun Unix OS 4.0.3.

* This work was supported by the Office of Health and Environmental Research Program of the Office of Energy Research, U.S. Department of Energy under Contract DE-AC03-76SF00098.

References

[1] V.M. Markowitz and A. Shoshani, "On the Correctness of Representing Extended Entity-Relationship Structures in the Relational Model", Proc. of *1989 SIGMOD Conference*, June 1989.

[2] V.M. Markowitz and A. Shoshani, "Name Assignment Techniques for Relational Schemas Representing Extended Entity-Relationship Structures", Proc. of *8th International Conference on Entity-Relationship Approach*, Toronto, 1989.

[3] V.M. Markowitz, "Representing Processes in the Extended Entity-Relationship Model", to appear in the Proc. of *6th International Conference on Data Engineering*, February 1990.

[4] V.M. Markowitz and F. Olken, "An Extended Entity-Relationship Schema for a Molecular Biology Laboratory Information Management System", Technical Report LBL-27042, May 1989.

[5] V.M. Markowitz and W. Fang, "*SDT* Programmer's Manual", Technical Report LBL-27843, November 1989.

Data Thesaurus for Physical Mapping

John L. McCarthy*
Information and Computing Sciences Division
Lawrence Berkeley Laboratory

Physical mapping of human chromosomes must integrate various types of data from different sources. One barrier to integration is that many pertinent biological entities (e.g., cell lines, DNA segments, genes, probes, restriction enzymes, etc.) are known by different names, abbreviations, and local laboratory codes. Multiplicity of names for the same entity (not to mention use of the same name for different entities) is a significant problem for both people and computer programs. Although a small, well-disciplined laboratory can address such problems by requiring that everyone use a single, standard local name for each entity, such an approach is unrealistic for a large laboratory that must interact with many external sources of information.

As an alternative to enforcing a single local naming standard, LBL's Human Genome Center plans to adapt a software mechanism called the data thesaurus to support systematic maintenance and automatic translation of synonyms and related information for major types of biological entities used in its laboratories. LBL's initial prototype data thesaurus will focus on entities pertaining to Chromosome 21, including genes, probes, restriction enzymes, and cell lines.

As demonstrated by its success in conjunction with another LBL scientific data project, the data thesaurus can serve a variety of purposes for both data administrators and end users. For data administrators, the data thesaurus will be a tool for maintaining, documenting and updating various controlled vocabularies. Names of genes, probes, cell lines, and so on will be registered as either primary terms or synonyms in the thesaurus before they can be used in other LBL databases. The thesaurus can then be used to provide lists of allowable values for such controlled entities and/or to validate relevant fields for either batch or interactive data entry.

In addition to automatic translation of synonyms, the thesaurus paradigm also can support automatic "explosion" of hierarchical groups or classes of entities. For example, if restriction enzymes are classed by both cutting frequency and vendor, people and programs could automatically access sets of enzyme names belonging to larger classes such as "infrequent cutters" or "Merck."

The data thesaurus requires standard database capabilities for access control, data integrity constraints, and so on, with special emphasis on retrieval of nested and repeating text data structures. Since other computer programs, as well as human users, will depend on it for name-based information, the thesaurus must be easily accessible to both via standard interfaces over local and wide-area networks. This combination of requirements may be difficult to achieve with a commercial relational data management system. We are currently trying to identify an appropriate data management system that we can use for the data thesaurus in conjunction with other Human Genome Project software.

*Bldg 50B - 3238, LBL, Berkeley 94720; Internet Email: JLMcCarthy@lbl.gov

Neural Net Applications to DNA Sequence Analysis *

McMullen, D. F., Lim, H.A., and Burnette, D.E.

Supercomputer Computations Research Institute

Florida State University

Tallahassee, FL 32306-4052

The main objective of this project is to develop neural network ("Connectionist") algorithms for performing several common operations in the analysis of DNA sequence data. Primarily these operations involve the comparison of a sequence with the contents of a data base or the alignment of a number of related DNA fragments. Current, nonconnectionist methods employ heuristic rules or a dynamic programming algorithm to define the degree of similarity desired. For long sequences or data with noise (point mutations or errors, or longer insertions or deletions) similarity searches and alignment can be performed more efficiently using a content-addressable memory scheme than with conventional methods. In order to test the efficacy of connectionist algorithms to the analysis of DNA sequence data, a simulated "chromosome" was constructed by overlaying a random sequence of bases with known sequence data associated with hc14 and regularly distributed features such as inverted palindromes and regions of known base concentration. The simulated chromosome ("simugen") is then "cleaved" by searching for restriction enzyme sites using a three layer back propagation network, and the fragments used in subsequent tests of alignment and data base search algorithms.

* Funded by DOE Energy Research through SCRI

Mapping Algorithms for the Probed Partial Digestion Problem

Abstract

Dalit Naor
Computer Science Division, U.C. Davis
and
ICS Division. Lawrence Berkeley Laboratory

The Probed Partial Digestion method partially digests the DNA with a restriction enzyme. A probe, known to be located between two RE cutting sites. is then hybridized to the partially digested DNA, and the sizes of fragments which the probe hybridizes to are measured. The objective is. then, to reconstruct the linear order of the RE cutting sites from the set of measured lengths.

A Backtracking algorithm, which runs in $O(2^{n-4})$ worst case. where n is the number of cutting sites, is given. The algorithm finds all possible orderings that are consistent with the data. It can be modified to handle inaccurate data. If the data set contains only the lengths but not their multiplicities (that is, band intensities are not taken into account) then the algorithm runs in $O(n^k 2^n)$, where k is the degree of multiplicity, which is believed to be small. A more general version of the problem that uses multiple probes is considered. Two special cases for which simplified solutions exist are pointed out.

The Backtracking algorithm has been implemented for the case where the data set is complete, and preliminary tests have shown that the stated worst case running time is over pessimistic.

Finally, we look at various deconvolution algebraic methods that were suggested in the literature to solve the Partial Digestion problem (when no probes are used). These methods are known to run in polynomial time and are based on algebraic algorithms for factoring polynomials; however, then are *only* applicable when the data is complete and accurate. We point out the difficulties in applying these methods to Probed Partial Digestion.

ROBUST METHODS FOR SIGNAL EXTRACTION AND CALIBRATION IN RESTRICTION FINGERPRINTS

David O. Nelson, Tom Slezak, Elbert W. Branscomb, and Anthony V. Carrano. Biomedical Sciences Division L–452, Lawrence Livermore National Laboratory, Livermore, CA 94550.

Analyzing restriction fingerprints for Chromosome 19 presents several difficulties not normally encountered in data generated, for example, for DNA sequencing. For instance, the data acquired is not in the typical "picket-fence" shape, but rather consists of a superposition of a random number of possibly overlapping peaks of varying sizes. In addition, the signal is corrupted by at least three distinct sources of noise: random noise that is uncorrelated from channel to channel, cross-talk from the other samples loaded in the same lane, but tagged with different colored dyes, and correlated noise corresponding to situations such as imperfections in the gel. We are developing robust, reliable methods for signal extraction and analysis in this complex environment.

The three major signal processing tasks consist of noise suppression, peak detection, and fragment size determination. We suppress the random, uncorrelated noise using a standard robust smoother ("4253H,twice"). We model the color bleed-through as a matrix equation $Ax = b$, where b is the observed data, A is an empirically-determined bleed-through transfer function, and x is the unknown signal we want. For each b, we color-correct by using Hanson's constrained least squares routine "SBOLS"[1] to minimize $\|b - Ax\|$, subject to $x \geq 0$. Enforcing non-negativity in the solution serves to ensure a physically meaningful result as well as to further suppress the noise in the signal. We have just begun to examine ways to suppress correlated "noise" due to gel imperfections and the like.

We are exploring two different, complementary ways to detect peaks. One way, based on smoothing splines, can quickly find peaks in parts of the signal where the structure is not too complex. The other approach, based on treating the peak detection problem as one of *deconvolution*, is about two orders of magnitude more computationally intensive, but can find peaks that simpler, more non-parametric methods cannot. In this model, we assume the given signal is a noisy convolution of a set of "impulse functions" (representing the peaks) by a blurring kernel (representing the diffusion process which occurs simultaneously with migration through the gel). If so, we can recover approximations to the impulse functions by deconvolving the signal with a representation of the blurring kernel. We are using Zhuang's basic approach to Maximum Entropy deconvolution.[2] However, instead of using his somewhat heuristic algorithm, we are solving his system of Differential-Algebraic Equations directly, using a state-of-the-art DAE solver called DASSL, developed at LLNL by L. Petzold.

In addition to the above noise suppression and data extraction tasks, we also must calibrate the peak locations to a standard. We determine fragment sizes from lane positions in two steps. First, we use a dynamic-programming algorithm which matches lengths from a known standard with peak positions corresponding to that standard (one standard is run in each lane). Finally, we interpolate intervening distances using Fritsch's monotone spline package. [3] [4]

---

[1] RJ Hanson (1982). Linear least squares with bounds and linear constraints. SNLA Report SAND82-1517.

[2] Zhuang et al (1987). IEEE Trans. Acoustics, Speech, and Signal Processing. ASSP-35:2, 208–218.

[3] FN Fritsch, RE Carlson (1980). Siam J.Numer.Anal. 17:2, 238-246.

[4] This work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

The Laboratory Notebook: A Relational Database for the Management of Physical Mapping Data

Debra Nelson, Carmella M. Rodriguez, and Thomas G. Marr
Los Alamos National Laboratory

We have designed and implemented a relational database to manage the data produced by the physical mapping effort for human chromosome 16 at Los Alamos National Laboratory. The database design is sufficiently general to be useful, with little modification, for most mapping strategies. The "Laboratory Notebook" database is meant to be an electronic version of the ubiquitous lab notebooks found in all molecular biology experimental facilities and was designed not only for management of data resulting from experiments, but also to manage information associated with materials, methods, and procedures. The database resides in Sybase, a relational database management system, on a Sun 4 computer operating under UNIX.

We have developed tools to support the flow of data from the laboratory into the database, and once in the database into various forms suitable for analysis and presentation. Electrophoresis gel image information (whole band report) including DNA fragment sizes are transferred directly to the database from the BioImage Visage 110, an image-processing workstation, where the photographs of stained gels are digitized and analyzed. Fingerprint annotation is added to the DNA fragments through a form-based user interface developed for data entry, editing, and retrieval. Reports from these data are created for direct input to contig construction programs (e.g., programs which estimate the probability of pairwise overlap).

Because the users of the "Laboratory Notebook" database are in most cases molecular biologists and not computer scientists, the user interface was designed to present a conceptual view of the data without burdening the users with the need to understand the structure of the database and details of data storage. The user interface was developed using the Sybase application APT-forms and requires very little training to use.

We are in the process of extending the implementation of the database and enhancing the user interface to allow on-line access to other data entities. We will continue to develop tools to facilitate automatic transfer and integration of our local data as well as those data being generated by our collaborators.

SIZE AND DNA BASE COMPOSITION ANALYSIS OF DNA FRAGMENTS.
D. Peters* and J. Gray. Lawrence Livermore National Laboratory, Livermore, CA

A dual laser, capillary electrophoresis apparatus has been assembled and used to classify DNA restriction fragments according to size and DNA base composition. In this system, DNA fragments stained with Hoechst 33258 (binds preferentially to AT rich DNA) and chromomycin A3 (binds preferentially to GC-rich DNA) are loaded electrophoretically into a 30 cm quartz capillary (50 μm id) filled with 1% agarose and separated electrophoretically by applying a 1KV potential across the capillary. The fragments pass through laser beams that are focussed through the capillary near its end. The lasers are adjusted to the UV and 442 nm to excite Ho and CA3, respectively. The fragments are classified according to their time of arrival at the laser beams (i.e., according to molecular weight) and according to the ratio of their HO and CA3 fluorescence intensities (i.e., according to DNA base composition).

Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract number W-7405-ENG-48.

SIMULATION OF PHYSICAL MAPPING

Karl Sirotkin and Eric Fairfield

Los Alamos National Laboratory

We have created a set of modular programs to evaluate strategies for physical mapping of the human genome. A person using these programs can: create a test genomic segment, generate clones from this segment, extract fingerprint data from each clone, test various strategies for determining pair-wise overlap between clones, reassemble the genome from these overlaps, display the resulting contigs, and evaluate the success of the reassembly.

The programs have been structured to evaluate many mapping strategies and to assemble contigs from real data as it becomes available. In anticipation of other users, both program and data files are readable by the user and the same symbolic parameter names are used throughout the data and program files. In addition, we have designed the program structure so that it is easy to tailor the installation for individual users.

These programs have been implemented in two phases. The first phase only used exact fingerprint data, while in the second phase there were controlled levels of error in the fingerprint data.

By using exact data, four different strategies reassembled similar percentages of a genome segment into contigs; although, the exact contigs were different. By combining information from all of these methods, the coverage of the genome by contigs increased. With errors in the fragment lengths, it is not yet clear whether particular strategies for reassembly of fragments into contigs make better use of the data. Small changes in the experimental errors seem to have large effects on contig generation.

The current set of programs have been delivered to a beta test site and will be available for testing at the conference.

HUMPTY: An Algorithm for Fully–Automated Contig Assembly

Tom Slezak, Elbert W. Branscomb, and Anthony V. Carrano. Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

Developing a physical map of a human chromosome in the form of an ordered library of cosmid clones involves assembling many thousands of cosmids. Manual contig assembly on this scale is neither practical nor necessary. We have developed an algorithm that automatically assembles contigs and determines a near–minimal spanning path with highly–confident determination of contig ends. Input to this algorithm is a sorted list of the LOD scores (odds ratio in favor of overlap) of all pair–wise combinations of DNA fragments. An optimized data structure is coupled with a depth–first greedy search strategy, yielding correct reconstruction of 8,000 simulated chromosome 19 cosmids in under 15 minutes on a Sun–4/260. The algorithm is coded in C for running under Unix. EcoR1 digests of selected chromosome 14 contigs have verified both contig membership and spanning path determination. Output from the HUMPTY algorithm can be viewed and manipulated graphically using the contig browser described in a separate poster by Mark Wagner of LLNL.

(This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W–7405–ENG–48.)

Special Requirements: Poster Presentation, no equipment needed

A Graphical Contig Browser Tool for DNA Mapping

Mark C. Wagner, Thomas R. Slezak, Elbert W. Branscomb, and Anthony V. Carrano. Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

We have developed a highly–automated system for the visualization of mapping information from cosmid clones. Input to this Browser tool is derived from HUMPTY, a program which takes the overlap probability for all pairwise combinations of cosmids based on fingerprint data and forms spanned sets of overlapping clones (contigs). This information is too complex and voluminous to be readily understood on paper. We developed a graphical contig browser to assist in comprehending and manipulating this data. Providing a color–coded visual representation of the data permits a better understanding of the relationships between the individual cosmids that comprise each contig. Our implementation is based on LLNL–developed graphics libraries that run on top of the industry–standard X11 graphics system (Sun and Stellar), and on Silicon Graphics Iris workstations. All coding is in C for use under Unix. Our prototype version is in daily use helping us to analyze the over 2,500 human chromosome cosmids fingerprinted and mapped to date. Work in progress will tie the contig browser to a Sybase relational database system, add the ability to view the raw cosmid fingerprint data from several cosmids simultaneously, and allow the ability to view and manipulate 2 contigs at once to allow gap closure from non–fingerprint data. This tool is our model for our planned graphical chromosome database browser that would allow ready access to all data generated on the Human Genome project.

(This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W–7405–ENG–48.)

Special Requirements: Poster Presentation, software can be demonstrated live if a color Sun–4 system is available.

Abstract for poster presentation to DOE Contractor's Workshop, Nov 3 - 4, 1989 in Santa Fe, NM, by Mr. John West, Principal Investigator, BioAutomation, Inc.

CUSTOM CHIP TECHNOLOGY FOR IMAGE SCANNING

Numerous applications exist in modern biology for image scanning. The push for automation in the Human Genome Initiative will expand these applications. We report on experimental work done in our laboratory in 1989, investigating the use of application specific integrated circuit (ASIC) technology to implement circuitry for image scanning. In our work we have used a 1.2 micron CMOS gate array. With isolated flip flop toggle rate capability of 100 MHz, we find that the balance of routing and logic resources on the chip, combined with the capacitance induced routing delays, allows implementations to date with up to 6 MHz operating frequencies. This is a frequency which is easy to interface with other components of a system. The parallelism possible in such a hardware implementation provides for high performance at this frequency. With moderate volume, the costs of such an implementation can also be quite attractive.

The Human Genome Information Resource

Scott L. Williams,[1] Rena Whiteson, David C. Torney, Robert D. Sutherland, Karen R. Schenk, Alice Sandstrom-Bertini,[1] Carmella M. Rodriguez, Robert M. Pecherer, Debra Nelson, Frances A. Martinez, Thomas G. Marr, James H. Jett,[2] C. Edgar Hildebrand,[2] Michael J. Cinkosky, and Christian Burks.[3] Theoretical Biology and Biophysics Group; T-10, MS K710; Los Alamos National Laboratory; Los Alamos, NM 87545; U.S.A.
[1]Computer Research Laboratory; New Mexico State University; Las Cruces, NM.
[2]Life Sciences Division; Los Alamos National Laboratory; Los Alamos, NM 87545.
[3]Corresponding contact: telephone, 505-667-6683; e-mail, cb%intron@lanl.gov.

The Human Genome Information Resource is focused on the need for better information management and analysis tools for physical mapping data (e.g., the data currently being generated in the context of the effort to generate a complete physical map of human chromosome 16 [Hildebrand et al. (1989) these proceedings]), and reflects a long-term interest in extending research to other, related data sets such as nucleotide sequences and genetic maps.

Because of the desirability of having "real" experimental data to examine and manipulate, and because of the close ties with the experimental group in LS-Division at LANL, the project has focussed initially on developing strategies and tools for supporting the flow of data from DNA gels into computers and, once in the computer, into various forms for analysis and presentation. This flow of data begins with the digitization and processing of electrophoretic gel images. Data corresponding to the clones analyzed on the electrophoretic gels are then passed into a computerized laboratory notebook [Nelson et al. (1989) these proceedings] based on a relational database management system. These data are made available for subsequent analysis of the clone fingerprints, leading to the development of contig maps [Torney (1989) these proceedings] and -- eventually -- comparison to other, related data sets that will allow for higher-order assemblies and ordering of contigs.

As one gets to the end of this flow path, the need for more sophisticated data management, analysis, and interface tools becomes evident. The thrust of the HGIR project will shift to the development of these tools, and their use to facilitate the cross-linking among multiple levels of physical mapping data, as well as between physical maps and other, related data sets (e.g., sequences and genetic maps). The design work on data structures for physical map and sequence data we are doing is being undertaken with this emphasis (and the future extension to yet unrecognized data structures) in mind. Finally, we plan to design an on-line system and set of interfaces allowing for the provision of these data and tools to a much more broadly-defined user community than the current in-house activity.

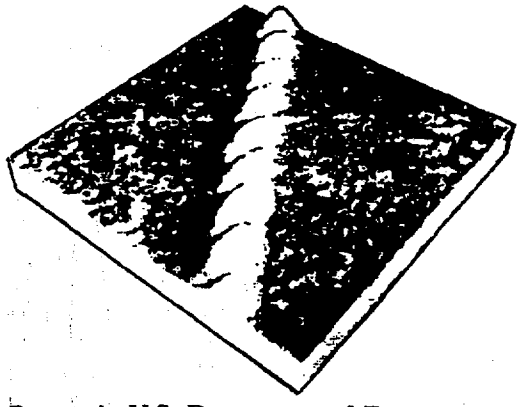
Scanning Tunneling Microscopy of Macromolecules

D. P. Allison, J. R. Thompson, K. B. Jacobson, R. J. Warmack, and T. L. Ferrell

Oak Ridge National Laboratory*
Oak Ridge, Tennessee

Beginning in 1982 when the first images were reported, the development of the scanning tunneling microscope (STM) has revolutionized microscopy in the 1980's. Although the primary applications are for research on metal and semiconductor surfaces, success in imaging a number of biological samples, mounted on a variety of conductive surfaces, have established STM as a valuable emerging technology in biological research. In 1987 we began our biological STM studies using tobacco mosaic virus (TMV), an easily identifiable rod shaped virus, to test the feasibility of using STM on naked biological samples. The STM operates by scanning a sharp tip a few atomic diameters away from a conductive surface and measuring changes in current or voltage as a function of changes in surface topography. We deposited TMV by spraying an aqeous solution of the virus on to evaporated or sputter-coated palladium-gold (Pd/Au) films supported on flat mica surfaces. Although TMV could be clearly identified its observed width was typically 70-100 nm instead of the known diameter of the virus, 18 nm. On evaporated Pd/Au substrates, tip traces revealed structures elevated above the substrate surface suggesting that the virus became coated with Pd/Au allowing normal conduction to occur. On sputter-coated substrates tip traces revealed depressed substructures. This is presumed due to the poor electrical conductivity of TMV causing the true differential tip motion to be recorded erroneously. Although not routinely obtainable, we have observed exceptional images of the virus revealing protein subunit structures separated by only 2.4 nm. We propose these images are the product of an exceptional tip, combined with an unusual conductivity of the virus.

In May of this year we obtained our first images of DNA. A circular plasmid of p BR 322 containing two genes for antibiotic resistance (tet^R and amp^R) was mounted on a graphite surface and imaged in air. A portion of one of these images is shown, clearly demonstrating the right-handed helix with an axial repeat spacing of 4.7 nm somewhat different than the 3.4 nm repeat expected of the B-form of DNA. With the routinely available resolution demonstrated in these images, the STM is currently capable of detecting conformational changes and binding of substances, such as repressor proteins, to DNA molecules. In the future, with improvements in methods for routinely attaching both single and double-stranded DNA to conducting substrate surfaces, we propose to use STM to identify nucleotide sequences in intact DNA molecules. This will be accomplished by hybridizing oligonucleotides of known sequences, suitably labeled for recognition by STM, to intact DNA molecules.



* Sponsored by the Office of Health and Environmental Research, U.S. Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

NOVEL DNA POLYMORPHIC SYSTEM: VARIABLE POLY A TRACT 3' TO ALU I REPETITIVE ELEMENTS.

S.E. Antonarakis, E.P. Economou, and A.W. Bergen.

Genetics Unit, Dept. of Pediatrics, The Johns Hopkins University School of Medicine, Baltimore, MD 21205.

DNA polymorphisms are extremely useful in the mapping of the human genome and the search for disease gene loci. Several categories of DNA polymorphisms include single nucleotide substitutions, variable number of tandem repeats, presence or absence of L1, Alu I elements or pseudogenes and variable number of dinucleotide repeats (CA)n or (CT)n.

We describe here a novel class of DNA polymorphisms: Variable number of adenylic acid residues (As) at the end of Alu I repetitive elements. To test the hypothesis that the poly A tract of Alu I sequences is polymorphic, three areas that contain the 3' end of an Alu I element have been selected for polymerase chain reaction (PCR) amplification. The first about 600 nt 5' to the ß globin gene (ß Alu) the second about 1100 nt 5' to the transcription initiation site of adenosine deaminase gene (ADA-Alu) and the third in IVS 1 of the factor VIII gene (F8 Alu). The PCR oligonucleotides were chosen from the "right arm" of the Alu I repetitive element and from a single copy sequence following the poly A tract and were used in a 1:10 concentration ratio respectively. The single copy oligonucleotides were end labelled with ^{32}P and the PCR product was electrophoresed in a 6% acrylamide sequencing gel. Due to a different length of the poly A tract in different alleles several polymorphic patterns were observed. Mendelian inheritance was demonstrated in CEPH families and nucleotide sequencing confirmed that the multiple allelism was due to the different number of As.

The ß Alu showed 32% heterozygosity for different (multiple) alleles in the grandparents and parents of the CEPH families. The ADA Alu showed 6 different alleles with frequencies of 29%, 10%, 28%, 27%, 7% and 1% in 267 independent chromosomes examined in the CEPH families. The observed heterozygosity for this polymorphism was 75% The F8 Alu showed no polymorphisms in 20 unrelated females from the CEPH families.

Since more than 10^5 Alu I sequences exist in the human genome their variable poly A tract (Alu-VpA) may prove to be one of the most abundant and useful polymorphic systems.

Imaging of DNA Molecules Deposited on Graphite. R. Balhorn*, M. Allen*, B. Tensch**, J.A. Mazrimas*, M. Balooch†, and W. Siekhaus†. *Biomedical Sciences Division, **Department of Applied Science, and †Chemistry and Materials Science, Lawrence Livermore National Laboratory, Livermore, CA 94550.

The conditions required for imaging DNA near atomic resolution with the scanning tunneling microscope are being examined as the first step in our effort to devise an alternate, electronic method for sequencing DNA. Biotinylated lambda phage DNA and a defined length synthetic duplex DNA have been deposited on highly oriented pyrolytic graphite (HOPG) and imaged in air by scanning tunneling microscopy. The lambda phage DNA was tagged with streptavidin coated 20nm gold spheres and the spheres located in the initial 4000Å X 4000Å scans. High resolution images of three attached strands of DNA show the variability in level of detail that can be observed. Helical coiling is detected in several regions, but it is obscured in others. The end of one molecule is unwound and the last 100-120 base-pairs have separated and are visible as single strands. Reproducible images of a 47 base-pair (bp) DNA sequence have also been obtained. Each molecule exhibits a similar periodic structure and length compatible with expected values. Interesting structural details are revealed, including the left handedness of the DNA helix in the G-C rich region, the presence of short, single-stranded "sticky" ends, and the major and minor grooves. This work was funded by the U.S. D.O.E. by the Lawrence Livermore National Laboratory under Contract W-7405-ENG-48.

TRANSPOSON Tn5 FACILITATED DNA SEQUENCING

D. E. Berg, S. H. Phadnis, T. Tomcsanyi, H.V. Huang and C. M. Berg*.
Depts. Molec. Micro., Washington Univ. Med. School, St. Louis, MO. 63110, and
*Cell & Molec. Biol., University of Conn., Storrs, CT. 06269

Bacterial transposons are being developed to place unique sites for DNA sequencing primers and multiplex probes throughout cloned DNAs, thereby minimizing the need for random DNA subcloning or repeated syntheses of new oligonucleotide primers. These experiments involve derivatives of Tn5, a transposon that inserts efficiently and quasi-randomly in diverse target DNAs.

Tn5supF. We constructed this 264 bp mini-transposon for insertion mutagenesis and sequencing of DNAs cloned in phage λ. Tn5supF is marked with the suppressor tRNA gene, *supF*. Insertions into amber mutant λ are selected by plaque formation on wild type *E. coli*. Insertions in non-amber λ are selected similarly using the *dnaB*-amber bacterial strain DK21 from David Kumit, a selection that exploits the need for DnaB protein during λ DNA replication.

Saturation mutational analyses of entire genomes will grow in importance as genome sequencing projects near completion. We have begun recombining Tn5supF inserts made in the λ/*E.coli* hybrid phage of Kohara et al. (Cell 50:495-508) into the *E. coli* chromosome. Because these phage are defective in λ repressor synthesis (cI^-) phage λ*b*221 *c*I857 was used as a co-infecting helper to supply repressor and thus permit survival of infected cells. Haploid bacterial recombinants were obtained readily with a *lacZ*::Tn5supF mutation. In contrast, only partial diploids (containing both mutant and wild type alleles) were obtained with *rpmD*::Tn5supF and *rpoA*::Tn5supF insertion mutations because *rpmD* and *rpoA* encode essential proteins. We conclude that Tn5supF-based reverse genetics, entailing first physical mapping and then phenotypic testing of new mutations, is well suited for analysing genes and sites found during DNA sequencing.

Deletion factory. Our recent experiments have shown that intramolecular Tn5 transposition generates deletions that place different regions of the target DNA close to a transposon end, analogous to exonuclease-based in vitro and IS1-based in vivo nested deletion strategies. For these tests we placed a synthetic Tn5 element next to a *sacB* (sucrose sensitivity) gene, so that deletions could be selected by sucrose-resistance. In contrast to results with IS1, the endpoints of deletions made by Tn5 transposition were widely distributed in target DNAs.

Supported by grants GM37138 and DE-FG02-89ER60862.

STRUCTURAL AND TRANSCRIPTIONAL ANALYSIS OF A CLONED HUMAN TELOMERE Jan-Fang Cheng+, Cassandra L. Smith* and Charles R. Cantor+, Departments of +Genetics and Development, *Microbiology, and *Psychiatry, College of Physicians and Surgeons, Columbia University, New York, NY 10032

Isolation of a human telomeric YAC clone, yHT1 (Nucleic Acid Res 17, 6109-6127), allows the characterization of a common DNA structure next to the telomeric TTAGGG repeats. Blot hybridizations using various portions of the yHT1 clone to probe against a panel of somatic hybrids indicate that this common subtelomeric structure spans at least 4 kb in length, and appears in multiple copies on some chromosomes but does not appear on the X chromosome. The complete DNA sequence of yHT1 has been determined. This clone contains an AT-rich region and a CpG island, separated by a human Alu repeat. Cross-hybridizations are weak in rodents when using various portions of the yHT1 clone as probes. However, the CpG island gives strong and distinct cross-hybridizing fragments. The significance of this evolutionarily conserved DNA sequence remains to be determined. Transcription patterns in this subtelomeric region have been examined by Northern blot analysis. Both polyA+ and polyA- transcripts were detected when probing with DNA isolated from the AT-rich region. Only polyA- RNA was detected when probing with DNA isolated from the CpG island.

MINIATURIZATION OF SEQUENCING BY HYBRIDIZATION (SBH): A NOVEL METHOD FOR GENOME SEQUENCING

Crkvenjakov, R., Drmanac, R., Strezoska Z., Labat I.,
Genetic Engineering Center, PO Box 794, 11000 Belgrade, Yugoslavia

Human genome sequencing based either on gel electrophoresis, or recently proposed hybridization (Drmanac et al. GENOMICS (1989) 4:114) methods requires automated equipment on macro scale and can not be imagined as a routine procedure. Macro scale is mandated due to the requirements of robotic positioning of samples on predetermined coordinates and polymer separation in gels. However, determination of oligonucleotide contents of DNA which underlies SBH theoretically allows the micro scale processes with micro separated samples altogether comprising a macro scale reaction. It is possible to use the determination which clone/probe is on which random micro position instead of placing clone/probe on predefined macro position or volume. We propose the use of micro discrete particles (DPs) as vehicles for samples/probes. The recognition of specific association of a DP and a clone/probe is achieveable by premarking of DPs and/or determining characteristics of clone/probe in situ. The most obvious ways of marking DPs are shape, size or color, or attaching to it a specific combination of known oligonucleotides. We offer two possibilities for human genome sequencing (Drmanac et al., manuscript in preparation). For direct SBH $1x10^7$ clones coming from 10 separate genome parts are bound to $1x10^3$ different DPs in as many macro reactions (or eventually in a single macro reaction). $1x10^3$ monolayers containing more than $1x10^7$ individual previously mixed DPs are each after DP identification hybridized with groups of 100 differently labeled octamers. To this end we have developed conditions for reliable short oligonucleotide hybridizations. For inverse SBH $1x10^{7-9}$ different DPs are prepared each carrying a unique 12-15mer and unique combination of 20 out of 40 marking oligos. No more than 5000 separate macro reactions are needed for their preparation. After 40 hybridizations with marker oligos to find association between 12-15mers and DPs in a monolayer, in 1-100 hybridizations with fragmented, end labeled human DNA data for sequencing are generated. The monolayer area covers at most 100 microscope slides. The data collection for both approaches needs automated image analysis giving speed of data bits acquisition of $1x10^6$/s. Finally a substantional computing has a major role to keep track of information and generate sequence (see accompanying abstract). The described miniaturization concept and ensuing savings make human genome sequencing immediately feasible in a laboratory pending technological development.

DNA fragment fingerprinting and/or sequence determination by Fourier analysis of coherent X-ray scattering

Joe W. Gray[1], James Trebes[2], James Brase[2], Daniel Pinkel[1], Thomas Yorkey[2], and Heinz-Ulrich Weier[1]
[1]Biomedical Sciences Division, [2]Laser Program
Lawrence Livermore National Laboratory, Livermore, CA 94550

This report describes the theoretical basis for rapid characterization of the distribution of labels (e.g., high Z scatterers such as Au, I or Br) along DNA molecules. In this approach, the DNA fragment to be characterized is amplified to ~10^7 copies (e.g., by in vitro DNA amplification), labeled (e.g., by hybridization to labeled oligonucleotides for fingerprinting or by incorporation of labeled bases during in vitro DNA amplification for DNA sequence analysis) and arranged as an array of straight (but not necessarily parallel) DNA molecules. The distribution of labels in the ensemble of linear DNA fragments is determined by Fourier analysis of the scattering pattern formed during irradiation with coherent X-rays. The Fourier analysis is simplified by labeling one end of the test DNA fragments with a distinct scatterer (e.g., by hybridization with a gold microsphere labeled oligonucleotide). Preliminary analyses suggest that sufficient scattering for DNA sequence analysis can be obtained from ~10^7 I-labeled DNA fragments with an exposure of a few minutes to existing 8 KeV x-ray sources synchrotrons, laser plasmas, electron beam sources. The key to success in this process is the creation of arrays of *straight* DNA molecules. Constrained electrophoresis in low ionic strength buffer is being investigated as a way of accomplishing this. Initial experiments are directed toward analysis of the locations of iodine labeled thymines in the DNA sequence:

CCC CCC CCC CCC CCC TAA AAA AAA AAT AAA AAT.

This work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract number W-7405-ENG-48.

High Resolution DNA Mapping by STEM

James F. Hainfeld
Biology Department
Brookhaven National Laboratory
Upton, NY 11973

A new method has been developed to map DNA and RNA such that specific sequences might be visualized in the electron microscope to within 3 to 5 base pairs.

Preliminary results have been obtained using the following test system. A 622 base pair (bp) sequence from pBR322 was excised with restriction enzymes and purified. Next, a 128 bp T7 piece was inserted at position 276 (giving 720 bp total). Equal quantities of the 622 bp and 720 bp fragments were denatured and renatured. This resulted in 50% formation of heteroduplexes (one 622 strand paired with a 720 strand) leaving the extra bases as a single stranded loop. A 26-mer oligonucleotide was synthesized that was complementary to a region of the single stranded insert. A chemical modification added a sulfhydryl at the 5' end of the oligo and the undecagold cluster (with a 0.8 nm diameter gold core) was covalently attached to it. Next, the oligo and heteroduplexes were mixed under renaturing conditions and examined in the Scanning Transmission Electron Microscope (STEM). Gold clusters were observed at the expected positions.

The gold cluster is about 10 Å from the base it labels (3 base pairs) and the accuracy of positioning a base from the end of DNA segments in the STEM is 2 bp, giving a total potential positional accuracy of 3-5 bp. This should prove useful in the physical mapping of genomes.

APPLICATION OF THE SCANNING TUNNELING MICROSCOPE TO STUDIES FOR DNA STRUCTURES

M. Salmeron, M. Bednarski, D.F. Ogletree, T. Wilson

Center for Advanced Materials
Materials and Chemical Sciences Division
Lawrence Berkeley Laboratory
Berkeley, California 94720

Scanning Tunneling Microscopy has great potential as a tool for the study of biological macromolecules, including DNA. The STM can be operated in air or in liquids, and image contrast does not depend on metal shadowing or replication methods. STM images of unshadowed DNA obtained in our laboratory and by other groups have demonstrated sub-nanometrer spatial resolution.

Initially it was believed that the STM would have limited application to biology since most interesting materials are non-conductors. Experiment has shown that the STM can image many molecules on conductive substrates that are insulators as bulk materials. Two major problems remain to be solved before the STM can be used as a routine tool for molecular biology.

The first problem is contrast - in the STM, contrast depends primarily on shape, so conductive and chemically inert substrates are required that are smooth on the sub-nanometer scale over areas of several microns.

The second major problem is fixation of molecules to the substrate. Tip-surface forces in STM are often sufficient to deform or displace molecules. Methods must be developed both to reduce tip-surface forces and to bond molecules to suitable substrates. We will show that in principle, this can be solved by constructing organic monolayers that possess reactive functional groups in the surface. Examples of such layers on boron doped Silicon substrates will be presented.

HUMAN REPETITIVE DNA SEQUENCES FOR USE AS MARKERS IN MAPPING THE HUMAN GENOME

C.W. Schmid, E.P. Leeflang, G. Wang

A 480 clone library of repetitive human DNA sequences is being analyzed to generate potential probes for use in mapping the human genome.

The library was screened for known repeated sequences and of the remaining 264 clones, 23 clones have thus far been selected for furthur study including lambda clone base sequence analysis, copy number, genomic arrangement and homology to rodent DNA.

A Probe-Based Mapping Strategy for DNA Sequencing with Mobile Primers

Linda D. Strausbaugh, Michael T. Bourke, Martin T. Sommer and Claire M. Berg
Department of Molecular and Cell Biology, The University of Connecticut, Storrs, CT 06269.

Currently popular large-scale methods for DNA sequence acquisition require sets of short, often random, DNA fragments adjacent to primer binding sites. An alternative sequencing strategy utilizes mobile transposons whose ends are used as primer binding sites, thus permitting large clones to be sequenced without fragmentation. We demonstrate a novel and efficient probe-based method for the localization and orientation of such transposon-borne primer sites, which requires no prior restriction enzyme mapping or knowledge of the cloned sequence. This approach, which eliminates the inefficiency inherent in totally random sequencing methods, is applicable to mapping insertions of any transposon in plasmids and will be particularly valuable for larger recombinant molecules in vectors such as cosmids and P1.

The transposons gamma delta (Tn *1000*) and Tn*5* show considerable promise for large scale sequencing. Although not as intensively developed as some other elements, gamma delta (Tn*1000*) inserts quite randomly, and plasmids containing a single gamma delta insertion can be obtained readily. A 6.7 kb *Eco*RI fragment of *Drosophila melanogaster* DNA cloned in pBR325 was chosen as the target because partial sequence analysis had shown that this fragment contains regions of atypical base composition. In this model system, we have used existing features of wild type gamma delta and this particular recombinant DNA: *Eco*RI cuts gamma delta in its control region and cuts the plasmid at the two plasmid-vector junctions. DNAs from plasmids containing gamma delta insertions were digested with *Eco*RI, and the resulting fragments electrophoresed, transferred, hybridized to radioactive probes, and visualized by autoradiography. Fifty insertions were easily mapped and oriented using one probe specific for an end of gamma delta and a second probe specific for an end of the cloned fragment.

Primers specific for unique subterminal segments at each end of gamma delta were used to prime dideoxy double stranded sequencing. Each transposon yielded at least 200 bp of sequence information from each primer. These results confirm the random nature of gamma delta insertion and demonstrate the effectiveness of probe mapping. Since transposition and resolution functions can be provided in trans, mini-gamma delta derivatives designed especially for probe mapping will be easy to construct.

Transposon-based probe mapping and sequencing bridges the gap between large cloned segments and unordered subclones. The "duplex" sequencing strategy described can be adapted to multiplex sequencing by inserting heterologous probe/primer sites in gamma delta derivatives.

Characterization and use of linking libraries in Chromosome 21 restraction map construction. Denan Wang, Akihiko.Saito, Jose P.Abad, William M.Michels, Cassandra L.Smith and Charles R.Cantor. Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720.

A top down approach to mapping and ultimately, sequencing the human genome starts by the construction of a low resolution restriction map of each chromosome. Effective procedures have been developed for constructing *Not* I linking libraries starting from chromosome-specific genomic libraries. Seventeen unique single copy *Not* I linking clones from human chromosome 21 were identified in two libraries. Their chromosomal origin was confirmed, and regional location established by using hybrid cell panels. Hybridization experiments with these probes revealed pairs of genomic *Not* I fragments and neighboring *Not* I sites. Additionally, partial digestion as well as cell line polymorphism strategies were use to see neighboring fragments. These strategies construct maps in regions for which linking clones are not identified.

Cloning of Yeast Artificial Chromosomes by electroporation. M. Bell and R. K. Mortimer, Lawrence Berkeley Laboratory and Department of Molecular and Cellular Biology, Division Of Biophysics, University of California, Berkeley, CA 94720. A strong bias against the cloning of larger Yeast Artificial Chromosomes (YACs) by spheroplast-PEG transformation has been observed. Although treatment with polyamines facilitates the cloning of larger YACs, the smaller Yeast Artificial Chromosomes in any given ligation mixture are cloned preferentially. In an attempt to eliminate this problem, we are exploring electroporation of both spheroplasted and intact yeast cells. Since electroporation of S. cerevisiae is a relatively new field, we are currently establishing basic electroporation transformation procedures with control plasmids.

References

Burgers, P. M. J. and K. J. Percival (1987) Analytical Biochemistry 163: 391-397.

Delorme, E. (1989) Applied and Environmental Microbiology 55: 2242-2246.

Mc Cormick M. K., Shero, J. H., Antonarakis, S. E and P. H. Hieter (1989) Technique, in press.

Poster Presentation

DETECTION OF DNA SEQUENCES WITH CHEMILUMINESCENCE

Irena Bronstein, Tropix, Inc., 47 Wiggins Ave., Bedford, MA 01730

Non-isotopic detection of DNA sequences has most commonly been achieved with fluorophores and, in some cases, with alkaline phosphatase as the label which can be detected with a colorimetric substrate BCIP/NBT. We have developed a new substrate for alkaline phosphatase 3-(2'-spiro-adamantane)-4-methoxy-4-(3"-phosphoryloxy) phenyl-1,2-dioxetane (AMPPD™), which chemiluminesces upon enzymatic dephosphorylation. This substrate when coupled with suitably engineered oligonucleotide probes provides ultrasensitive detection of DNA in Southern blots, and rapid detection of DNA sequences in the genomic sequencing protocols. DNA probes which were labeled with biotin and incubated with streptavidin-alkaline phosphatase and AMPPD allowed the detection of subpicogram quantities of target DNA using Southern analysis. Chemiluminescent detection of DNA sequences using the genomic sequencing procedure of Church and Gilbert revealed images of sequence ladders on x-ray film with exposure times of less than 30 minutes, as compared to 40 hours for a similar exposure with a ^{32}P labeled oligomer. The demonstrated shorter exposure times would permit more frequent serial reprobings of DNA sequences.

The Separation of Non-Denatured DNA Fragments with Electrophoresis Gels that are Easily Formed by Crosslinking an Acrylamide-Rich Copolymer

A novel way of forming electrophoresis gels and separations that can be achieved with these gels are described. The gel-forming procedure is straightforward, begins with a stock solution of copolymer and does not involve toxic chemicals, oxygen exclusion, free radical polymerization or heating. Gels of polymer content greater than 2% are readily obtained and these provide media in which excellent separation and resolution of non-denatured DNA fragments up to 5,000 bp can be achieved.

Authors: Kenneth G. Christy, Jr., David B. LaTart, Hans W. Osterhoudt and Ignazio S. Ponticello. Life Sciences Research Laboratories, Eastman Kodak Company, Rochester, New York 14650-2122.

HWO:jrs/#081C
10/13/89

Instrumentation Development for Molecular Biology

J.B. Davidson
Instrumentation and Controls Division
Oak Ridge National Laboratory

Ultra low light level detection and imaging are techniques with potentially wide application in several areas of molecular biology. Among these are analysis of 2-D protein gels, sequencing gels and mapping blots. In addition, they can provide the basis for 2-D detectors in the important area of 3-D molecular structure determination by neutron and x-ray crystallography and small angle scattering. In theses applications, film can be eliminated and images of the fluorescent and radioactive bands and diffraction spots can be accumulated in a digital memory for analysis. The principles can be applied at the microscopic level for neuronography and in situ hybridization studies.

Progress will be reported on:

- -Electronic autofluorography developments
- -A "lensless" radiation microscope development
- -A P.C. based 2-D detection system for neutron and x-ray diffraction
- -A novel viewing aid to reading sequencing films, the "Unsmiler" (Demonstration)

Advanced Concepts for Base Sequencing in DNA

J. H. Jett, R. A. Keller, J. C. Martin, E. B. Shera
Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
(505) 667-3843, FTS 843-3843

We are addressing the problem of rapidly sequencing the bases in large fragments of DNA. The ideas presented represent the combined effort of a multidisciplinary task force composed of physicists, physical chemists, cellular and molecular biologists, and organic chemists. To reduce mapping requirements, the emphasis is on sequencing methods that are rapid, require little DNA, and are capable of sequencing large fragments. After evaluation of several physical approaches to sequencing, the decision was made to proceed with a modified-flow cytometer approach that employs laser-induced fluorescence to detect individual fluorescent molecules. A large fragment of DNA, approximately 40 kb in length, will be labeled with base-identifying tags and suspended in the flow stream of a flow cytometer capable of single molecule detection. The tagged bases will be sequentially cleaved from the single fragment and identified as the liberated tag passes through the laser beam. We are projecting a sequencing rate of 100 to 1000 bases per s on DNA strands approximately 40 kb in length.

Experimental Apparatus for Pulsed Field Electrophoresis Research

W.F.Kolbe, J. E. Katz, S. E. Lewis and J. M. Jaklevic

Engineering Division and Human Genome Center
LBL etc.etc.

A test apparatus for research and development involving pulsed-field gel-electrophoresis has been constructed. The system includes a 24-node computer-controlled power supply capable of independent programming and sequencing of individual attached electrodes. An experimental gel-box associated with the system is designed to provide flexibility in terms of individual electrode design and array geometries. A closed loop cooling system maintains precise control of buffer temperature with minimum perturbation to the streamline flow across the gel. Computer calculations are used to define the electrode voltage distributions required to generate specific elctric field profiles within the gel-region. A repetitive cycle of the calculated voltage distribitions can be imposed on the electrode array under software control in order to implement a variety of existing and experimental pulsed-field electrophoresis protocols. A three dimensional precision manipulator allows computer controlled scanning of the gel with interchangeable probes in order to map the spatial distribution of voltage, buffer temperature, pH, etc. Details of the system will be described and preliminary results obtained using a linear variation in the dwell times of homogeneous, switched electric fields will be presented.

CLONING LARGE HUMAN INSERTS

Andrew A. Kumamoto, Ronald Law*, Robert Deans+, and Philip Youderian. California Institute of Biological Research, La Jolla, CA 92037; *MBI, UCLA, Los Angeles, CA 90024; +Department of Microbiology, USC School of Medicine, Los An · les, CA 90033.

We have devised a protocol for preparing large fragments of human DNA (250-450 kbp) based on a procedure for preparing phage P22 DNA. This procedure is rapid (about three hours), requires a minimum number of manipulations, and, most importantly, does not use phenol or lengthy dialysis steps. 44 kbp phage P22 DNA prepared is this manner has been shown to have 4 to 5-fold greater biological activity than phenol-extracted phage DNA.

We are developing a novel vector that will permit the cloning of these large (250-450 kbp) segments of human DNA. This vector will carry large human inserts as segments of E. coli F plasmids. Several human (and, for that matter, E. coli) sequences cannot be maintained on high copy number vectors, including traditional cosmid cloning vectors, and only can be recovered intact as single copy clones. Therefore, we are testing three different single-copy origins derived from F (oriV, oriVII, and oriS) as possible vector origins. To facilitate the preparation of large amounts of insert DNA, these vectors will be modelled after F'::Mud-P22 elements that may be amplified to very high copy number (e.g., 500 copies/cell of a 250 kbp plasmid) after a terminal induction event.

MANIPULATION OF SINGLE DNA MOLECULES IN A MICROSCOPE STAGE. M. F. Maestre‡, M. Miller‡, S. Goolsby‡, W. Johnston†, C. Bustamante*, Lawrence Berkeley Laboratory, Berkeley CA. and *Chemistry Dept. University of New Mexico. Supported by NIH Grant #AI08427 and *GM32543 and by the † Director, Office of Energy Research, Scientific Computing Staff, U.S. Dept. of Energy, contract DE-AC03-76SF00098, and by the ‡ † LBL Human Genome Center.

The aim of the project is to develop techniques that will permit the visualization and manipulation of a selected single DNA molecule for the purposes of mechanical or chemical alteration of the molecule. To this goal we have designed a microelectrode chamber with a spatially distributed electrode network of microscopic dimensions to be used in l with a fluorescent imaging. The electrode network consist of 24 electrodes of dimensions of 10 micrometers in thickness separated from each other by 10 micrometers. With the development of the technique, we hope to do controlled single molecule chemistry. By this we mean the ability to bring a DNA molecule to a particular enzyme which is immobilized in a specific region of the microchamber and after reaction to move the resultant products to further manipulation or reaction in other regions of the chamber.

The project can be divided into approximately four parts: a) design and operation of the microelectrode array, b) determination of the field strengths in the spaces in the microelectrode array c) manipulation strategies for the orientation, stretching, immobilizing and selection of single DNA molecule in the microscopic stage by the use of the forces induced by the microelectrode array. d) Measurement of the dynamics of electro-optic relaxation, field free relaxation and the viscoelastic properties of single DNA molecules. This will permit the measurement of the optical properties of single oriented DNA at the microscopic level. It allows the observation of alterations of local structure (at the resolution of the microscope objective) monitored through the changes in the polarization parameters or intensity of the fluorescent signals.

The measurement consist of labeling the DNA molecule with a fluorescent intercalating dye, e.g. acridine orange (AO), or ethidium bromide etc. The labeled DNA molecule is then placed on a slide that has micro-electrodes deposited on its surface. The position, motion and shape of the molecule is visualized by the technique of epi-fluorescence microscopy. The fluorescence emitted by the labeled DNA is visualized by and intensified video camera. The voltage of each of the 24 electrodes is controlled by a computer, allowing the manipulation of the DNA on a microscopic scale. The images are continuously recorded for processing by image analysis programs.

As an example of the possible measurements available with this technique, we have studied the field-free relaxation behavior of a single DNA molecule that has been extended by an electric field and allowed to relax at zero field strength under Brownian motion. The length of the stretched DNA is measured as it shortens as a function of time. The relaxation time is a measure of the hydrodynamic forces affecting the DNA molecule, the elasticity of the molecule, and the total length of the DNA (or molecular weight).

When the DNA molecule is totally stretched under the applied electric fields, waves can be seen propagating down the length of the molecule. These waves are analogous to the vibrations seen in a stretched string and are a function of the tension of the string and the modulus of the material. Consequently, the measurement of the velocity of propagation of the waves will give information on the rigidity of the DNA molecule (i.e. the Youngs modulus), a quantity heretofore inaccessible to direct measurement.

Title: **Image Acquisition and Processing System for the Analysis of Fluorescence from Stained DNA Gels,** Ronald A. McKean

Current methods for analyzing DNA in gels using fluorescence techniques are inadequate since results cannot be easily accessed by a computer and are difficult to reproduce. The lack of instrumentation for converting a fluorescing image into a digitized record for computer entry and the lack of techniques for standardizing analysis performed under varying conditions, severely limit usefulness of DNA separation in gels. Overcoming these problems is essential as needs increase for efficient DNA analysis.

KMS Fusion, Inc. is in the second phase of a DOE sponsored SBIR to develop a system for direct analysis of fluorescence from stained DNA gels. The system development consists of a versatile optical scanner, control and analysis software, and acquisition and control interfaces to an IBM-AT compatible personal computer.

The scanner employs a unique optical system capable of high spatial and photometric resolution, as well as low light level imaging. Scanning techniques are used to image stained DNA directly from agarose or polyacrylamide gels. The scanner incorporates many features, including optical/sensor calibration, modular filter and excitation designs, and selectable image apertures and magnifications. It is operational through front panel controls or an RS232 port. Image data is made immediately available to the computer for further processing.

The software controls the scan procedure, processes image data, creates compact data files, and presents results in a graphical manner. Lane data can be readily standardized to provide results in units of concentration and molecular weight. Statistical features are also provided that allow direct comparison of results from lanes contained in one or more gels.

The system utilizes the popular, and inexpensive, IBM-AT compatible personal computer. Data acquisition electronics and RS232 interface are placed within the computer.

This system offers a practical, inexpensive solution to problems long associated with gel analysis of DNA. Direct quantitation of fluorescence in stained DNA gels allows immediate access to the data by the computer. Computer analysis and processing allow data to be presented in units of concentration and molecular weight. Results from gels electrophoresced under varying conditions can be compared directly. The analyses are presented graphically as plots or reconstructed gel images. Unique data processing techniques allow gel data to be archived using minimal memory.

ESTIMATION OF THE DNA CONTENT OF HETEROMORPHIC AND ABERRANT CHROMOSOMES BY BIVARIATE FLOW KARYOTYPING. Barb Trask*, Ger van den Engh, Joe Gray; Lawrence Livermore National Laboratory, Livermore, CA

For flow karyotyping, chromosomes are analyzed for DNA content and relative base composition on a dual beam flow cytometer. Improvements in sample preparation, instrument accuracy and analysis software allow discrimination of all human chromosomes except 9-12. We have determined the relationship between peak location in a flow karyotype and chromosomal DNA content determined by quantitative microscopy (CYDAC). Maternal and paternal-derived homologs of many chromosomes can be distinguished on the basis of small differences in DNA content (3-5%). Heteromorphism in a population of normal donors was studied. The chromosomes showing the most variation are Y,21,22,13,14,15,16, and 9. The least heteromorphic chromosomes are X,2,4,7,8, and 17. Some variants could be correlated with variation in the size of regions identified by chromosome-specific repetitive sequence probes. DNA contents determined from flow measurements of hetermorphic chromosomes are correlated closely to earlier CYDAC measurements on the same individuals. Family studies show that heteromorphisms are faithfully inherited. Deletion and insertion detection using flow karyotyping will be discussed in light of normal heteromorphism. For example, the DNA content of chromosome 21 can differ by as much as 50% among normal individuals. A series of lines with X chromosome abnormalities with DNA contents ranging from 0.49 to 1.85 times that of a normal X was flow karyotyped. Measured DNA content change was linearly related to that predicted by cytogenetics. Small deletions in chromosome X of ≈2 Mbp, below the limit of banding resolution, were detected and quantified using flow karyotyping. Flow karyotyping is also a means to rapidly monitor somatic cell hybrids for the presence of intact human chromosomes.

Work performed under the auspices of the U.S. DOE (contract W-7405-ENG-48) with support from USPHS grant HD-17665.

ABSTRACT TITLE:

Chemiluminescent Imaging of DNA in Electrophoretic Agarose Gels

AUTHORS:

Doris Willis, B.S. Medical Technology,
Paul A. Gray, M.S. Biology,
Rosemarie F. Werba, M.S. Environmental Education,
Robert W. Coughlin, Ph.D. Chemical Engineering, P.E.,
Edward M. Davis, Ph.D. Biochemistry, SymBiotech, Inc.,
8 Fairfield Boulevard, Wallingford, CT 06492,
(203) 284-7465.

ABSTRACT:

We have developed a new chemiluminescent (C.L.) labeling procedure for visualizing DNA in electrophoretic (E.P.) gels that is safe and sensitive. C.L. imaging eliminates the danger of mutagenic dyes such as ethidium bromide (EtBr) and hazardous u.v. light. In addition, C.L. imaging of DNA in electrophoretic gels is more sensitive than EtBr staining. When Lambda DNA in agarose E.P. gels is stained in EtBr (5 ug/ml) for 30 minutes and then destained in 50 mM Tris buffer, pH 6.5, for 30 minutes only 3 ng of DNA is detectable. In contrast, 0.8 ng of DNA is detectable by C.L. imaging. Our procedure employs streptavidin-horseradish peroxidase (SA-HRP), luminol, and peroxide. DNA is biotinylated with photo-active biotin (Photoprobe, Vector Laboratories), electrophoresed in 1% agarose in TBE buffer at 10 v/cm for approximately one hour, affinity labeled with SA-HRP, and soaked in luminol-peroxide solution for 5 minutes. The luminous image of DNA in the E.P. gels is recorded with Polaroid 612 film using contact prints. Both Hind III Lambda DNA and KB DNA ladders have been visualized with this technique.